Spectral Regularization for Max-margin Sequence Tagging

Ariadna Quattoni · · · Universitat Politècnica de Catalunya⋆
Borja Balle · · · McGill University
Xavier Carreras · · · Universitat Politècnica de Catalunya⋆
Amir Globerson · · · The Hebrew University of Jerusalem

(⋆) now at Xerox Research Centre Europe

Abstract

- Max-margin learning of latent-variable sequence predictors
- Function class: Observable Operator Models
- Contributions:
  - ⋆ Max-margin completion of a Hankel matrix
  - ⋆ Spectral regularization for structured prediction
  - ⋆ Learning formulated as convex optimization

Sequence Tagging

- Task: map input sequences to output sequences

  h i p p o p o t a m u s
  h I p - x p a t x m x s
- Task Loss $\ell(\cdot,\cdot)$: Hamming distance
- Formulation using a scoring function

$$F : (\mathcal{X}\times\mathcal{Y})^\star \to \mathbb{R}$$

$$\hat{y}(x) = \operatorname*{argmax}_{y\in\mathcal{Y}^T} F(x,y)$$

- Max-margin Structured Prediction: given m training examples and a class of functions $\mathfrak{F}$ solve

$$\operatorname*{argmin}_{F\in\mathfrak{F}} \sum_{i=1}^{m} L(x^i,y^i;F) + \tau R(F)$$

  - $L(x,y;F)$ is the structured hinge loss

$$L(x,y;F) = \max_z \left[F(x,z) - F(x,y) + \ell(y,z)\right]$$

  - $R(F)$ is a regularization penalty for $\mathfrak{F}$

Usual Function Classes

- Factorized Linear Models of order k (e.g. CRF)

$$F(x,y) = \sum_{t=k+1}^{T} w\cdot\phi(x,y_{t-k:t})$$

$\longrightarrow$ *tractable but very dependent on* ϕ $\longleftarrow$

- Latent-variable Models of order k (e.g. latent SVM, HCRF)

$$S(x,y,h) = \sum_{t=k+1}^{T} w\cdot\phi(x,y_{t-k:t},h_{t-k:t})$$

$$F(x,y) = \sum_h \exp\{S(x,y,h)\}$$

$\longrightarrow$ *intractable prediction and learning* $\longleftarrow$

IO-OOM: Input-Output Observable Operator Models

Definition

$A = \langle \mathcal{X},\mathcal{Y},n,\alpha,\{A_{a,b}\},\beta\rangle$

- Alphabets: input $\mathcal{X}$, output $\mathcal{Y}$
- n states (i.e., hidden dimensions)
- Initial weights $\alpha\in\mathbb{R}^n$
- Final weights $\beta\in\mathbb{R}^n$
- Operator for each bi-symbol

  $A_{\sigma,\delta}\in\mathbb{R}^{n\times n}\quad \sigma\in\mathcal{X},\delta\in\mathcal{Y}$

Learning IO-OOM

- Regularizer: number of states $n = |A|$
- Piecewise objective

$$\operatorname*{argmin}_{A\in\mathfrak{F}} \sum_{i=1}^{m} L(x^i,y^i;F_k) + \tau|A|$$

$\longrightarrow$ *non-convex* $\longleftarrow$

Prediction with IO-OOM

- Scoring function using h

$$S(x,y,h) = \alpha(h_0)\left(\prod_{t=1}^{T} A_{x_t,y_t}(h_{t-1},h_t)\right)\beta(h_T)$$

- Scoring function marginalizing h

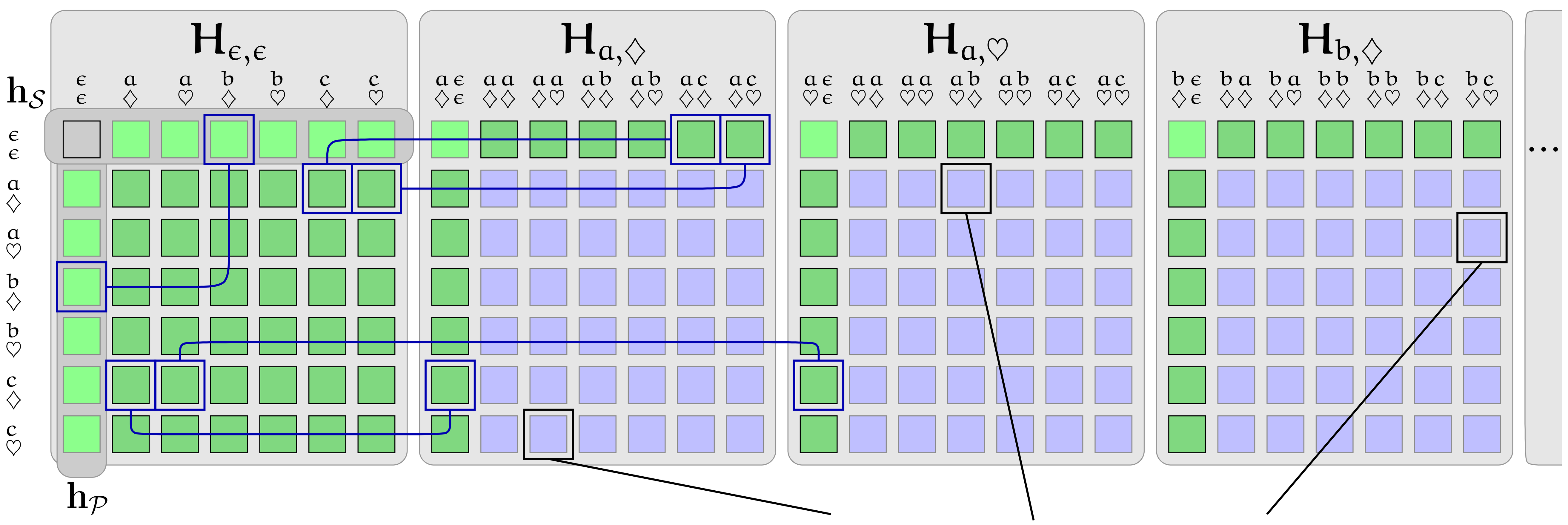
$$F(x,y) = \sum_h S(x,y,h) = \alpha^\top A_{x_1,y_1}\cdots A_{x_T,y_T}\beta$$

- Global piecewise prediction of order k

$$\hat{y}_k(x) = \operatorname*{argmax}_y \sum_{t=k+1}^{T} F(x_{t-k:t},y_{t-k:t}) = \operatorname*{argmax}_y F_k(x,y)$$

The Hankel Matrix of IO-OOM

$\mathcal{X} = \{a,b,c\}\quad \mathcal{Y} = \{\diamondsuit,\heartsuit\}$



$$F_{k=3}\left(\begin{smallmatrix}c&a&a&b&c\\ \heartsuit&\diamondsuit&\heartsuit&\diamondsuit&\heartsuit\end{smallmatrix}\right) = F\left(\begin{smallmatrix}c&a&a\\ \heartsuit&\diamondsuit&\heartsuit\end{smallmatrix}\right) + F\left(\begin{smallmatrix}a&a&b\\ \diamondsuit&\heartsuit&\diamondsuit\end{smallmatrix}\right) + F\left(\begin{smallmatrix}a&b&c\\ \heartsuit&\diamondsuit&\heartsuit\end{smallmatrix}\right)$$

- Definition: $\mathbf{H}\in\mathbb{R}^{\mathcal{P}\times\mathcal{S}}$ for A, using prefix-suffix sets $(\mathcal{P},\mathcal{S})$ as basis, with: $\mathbf{H}((u,w),(v,z)) = F(uv,wz)$
- The Fundamental Theorem of Weighted Automata (Schützenberger 1961; Carlyle & Paz 1971; Fliess 1974):

  F computed by A with n states $\iff$ $\operatorname{rank}(\mathbf{H})\le n$ for any basis $(\mathcal{P},\mathcal{S})$
- Main advantage: piecewise objective L is convex with respect to H
- Method:
  1. Obtain $\mathbf{H}$ optimizing L, via matrix completion techniques (Balle & Mohri 2012)
  2. Recover A from $\mathbf{H}$ using the spectral method of (Hsu, Kakade & Zhang 2009)

     2a. Take the reduced SVD of $H_{\epsilon,\epsilon} = U\Sigma V^\top$

     2b. $A_{\sigma,\delta} = (H_{\epsilon,\epsilon}V)^+H_{\sigma,\delta}V$; $\alpha^\top = h_{\mathcal{P}}^\top V$; $\beta = (H_{\epsilon,\epsilon}V)^+h_{\mathcal{S}}$

Convex Optimization

- Objective using Hankel and rank

$$\operatorname*{argmin}_{H\in\mathbb{H}(\mathcal{P},\mathcal{S})} \sum_{i=1}^{m} L(x^i,y^i;\mathbf{H}) + \tau\operatorname{rank}(\mathbf{H})$$

$\longrightarrow$ *non-convex* $\longleftarrow$

- Objective using Hankel and nuclear-norm

$$\hat{H}_S \in \operatorname*{argmin}_{H\in\mathbb{H}(\mathcal{P},\mathcal{S})} \sum_{i=1}^{m} L(x^i,y^i;\mathbf{H}) + \tau\|\mathbf{H}\|_*$$

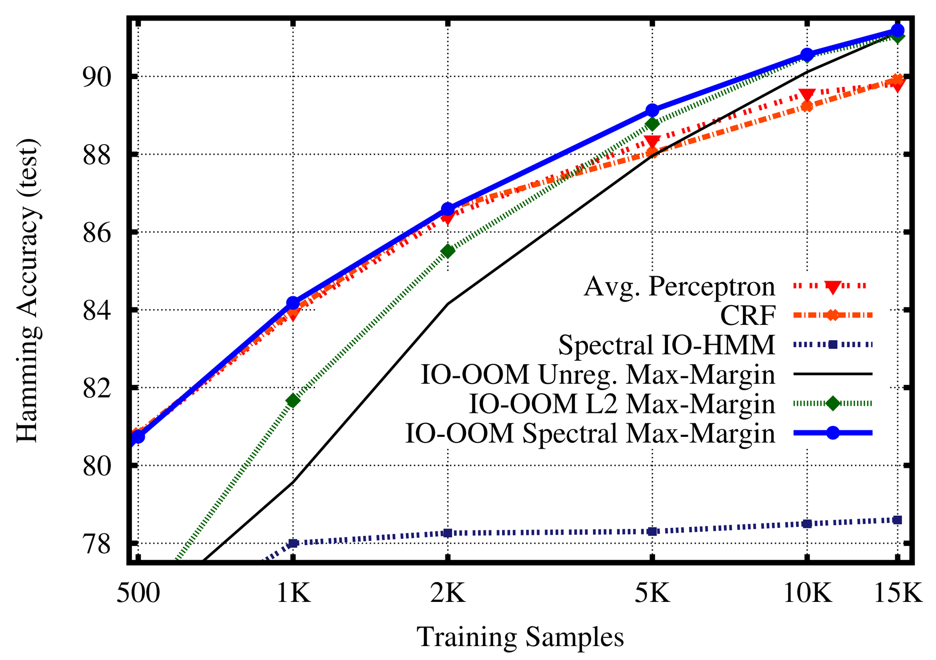
$\longrightarrow$ *convex* $\longleftarrow$

- We use Forward Backward Splitting (FOBOS) (Duchi & Singer 2011), based on gradient steps and proximal operators.

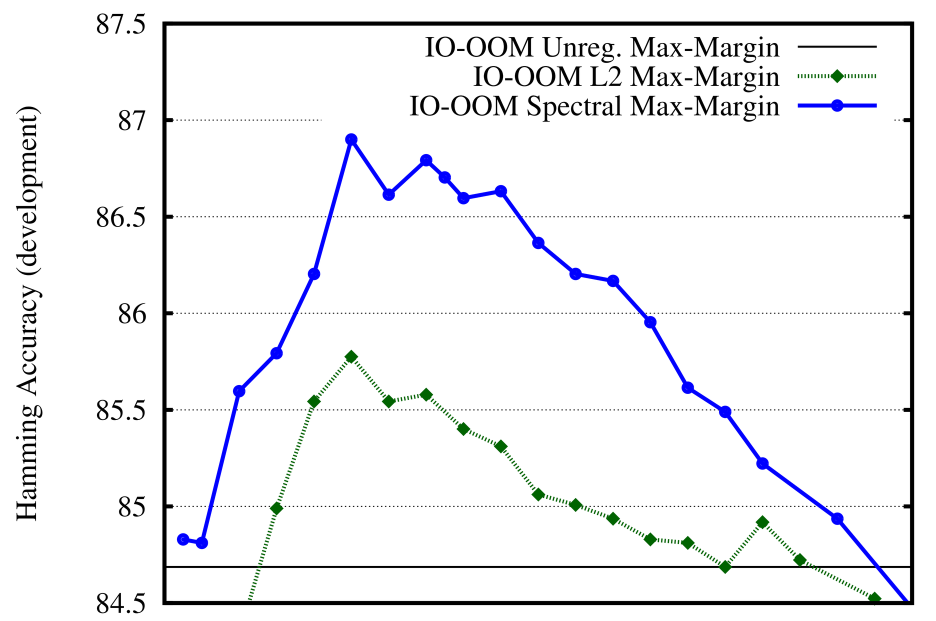
Experiments

- Phonetic transcription, "Nettalk" data ($|\mathcal{X}| = 26$, $|\mathcal{Y}| = 51$) (Sejnowski & Rosenberg, 1987)
- Methods compared:
  - IO-OOM Max-margin with
    - Spectral regularization
    - L2 regularization
    - no regularization
  - Factorized linear model (CRF, avg. perceptron)
  - IO-HMM trained directly with spectral method

Learning Curve



Regularization Path



Acknowledgements: This work was supported by projects XLike (FP7-288342), ERA-Net CHISTERA VISEN, TACARDI (TIN2012-38523-C02-00) and by the ISF Centers of Excellence (grant 1789/11). X. Carreras was supported by the Ramón y Cajal program of the Spanish Government (RYC-2008-02223). B. Balle received support from NSERC and the James McGill Research Fund. Xerox Research Centre Europe (XRCE) provided travel support for A. Quattoni and X. Carreras.